Formula One Driver Skill level Prediction

1st Mohit Gaggar
PES University
Bangalore, India
gaggarmohit@gmail.com

2st Ishan Agarwal
PES University
Bangalore, India
ishanagarwal1605gmail.com

3st Akash Kumar Rao
PES University
Bangalore, India
akashkrao99@gmail.com

***Abstract*—Formula One being a very competitive and high revenue generating sport, the scope of applying data analytics to maximising revenue is a common procedure. One such major task is to find out how good a driver actually is. All teams in formula one do analytics to find the best drivers in current times to generate their hiring options for the forthcoming season. In this project we try to separate team and car effects from the driver skill level and give rankings to drivers based on pure talent.**

***Index Terms*—formula one, analytics, driver skill level, ranking, increase revenue**

I. Introduction

Formula One is the topmost level of motorsport racing where multiple constructors develop one of a kind racing cars driven by the best drivers around the world. These constructors compete against each other in around 20 races every year, with each race being held in a different country, on different tracks. Top tier constructors in formula one spend million of dollars on car research and development.These cars can reach top speeds of above 360 Km/hr and can corner at 200 km/hr.Winning races can be considered a direct relation to developing the best cars and finding the perfect driver for the car who can push these cars to their limits and win races.The motivation to win races for teams is brand name publicity, huge amount of prize money, new sponsors who can invest in their team. Formula one uses a strong points system to track wins and positions of drivers and constructors throughout the season. According to this the winning driver of each race gets 25 points, second place gets 18, third 15 and keeps decreasing till the 10th place which fetches the driver 1 point.All other drivers get 0 points.The constructor points are calculated by adding points scored by the drivers driving for that constructor. In the end of a season the driver with the maximum points is declared formula one world champion, the constructor with maximum points as the winning constructor. In this project we use data collected from races from 1950s up till the current 2020 season and calculate driver rankings for every decade. Driver skill can be measured by considering the race results and devising how much of the race result was due to pure driver talent.

II. Related Work

A. Uncovering Formula One driver performances from 1950 to 2013 by adjusting for team and competition effects

Paper providing driver rankings within and between eras focussing on the driver scores instead of solely basing on finishing positions.This is an important takeaway as scores are better representations of a particular driver's performance in that season as compared to finishing position.

They have properly identified the different factors accountable for a driver's performance including the team performance some factors like pit stop times and performance of cars;competition with other drivers which affect scores and difficulty of scoring in that particular season .They have also taken into consideration faults which are again classified into driver and non driver issues which contribute to final scores .

They have used a simple linear statistical model for the predictor which takes into account all the aforementioned factors.

The limitation of the paper is that the classification of driver and non driver failures do not account for some cases for example where the driver may have caused the mechanical fault or how a vehicles poor handling may have caused the driver to make mistakes ,etc.The paper also doesn't account for how driver performance may be affected by different handling styles suiting them.

B. Who Is The Best Formula 1 Driver? An Economic Approach to Evaluating Talent

The paper aims at answering the question "Who is the best Formula 1 driver of all time". To answer this question it tries to separate the talent of drivers from the performance of their cars.

Despite Michael Schumacher having the most number of wins, the paper concludes that Juan Manuel Fangio is the best driver of all time, followed by Jim Clark using records from all the drivers' past races.

This factor of comparatively stronger opponents is also considered by the author and is found out to be an important factor

They consider finishing position as the dependent variable and control for team-year using dummy variable in a standard OLS-estimate single-level regression and also control for other variables in the simple model.

This causes the model to have too many variable and the simple nature of the model does not provide the best results possible. Another problem this comes across is that it takes into account the finishing positions instead of scores,the finishing positions tend to be a less accurate measure of a driver's performance compared to scores.

C. Formula for success: Multilevel modelling of Formula One Driver and Constructor performance, 1950-2014

This paper aims to study the drivers through the years and gives an overall ranking.

They have used a complex multilevel logistic regression model that tries and includes variables like teams of the drivers, points scored by drivers in each race, weather conditions to name a few.

They have learnt from the above 2 papers to propose this multilevel technique to compare drivers, they also consider complex analysis techniques like failure rate of one driver and how that affects his teammate.

The issue with the paper is that they include too many variables and over analyse data.

This means that few small details which should have been very important are not taken to be as important due to all the excess variables

III. Data

Data we have chosen is the Formula one dataset taken from Ergast Developer website (link http://ergast.com/mrd/db/) The data consisted of 13 tables namely circuits (stores information about the cicuits), constructorResults(stores the results achieved by each constructor after every race), constructorStandings(stores the season wide ranking of each constructor after every race), constructors(stores information about the consturctor like nationality, Id), driverStandings(stores overall standinging of the driver season wide after each race), drivers(stored driver information like Id, nationality, date of birth), lapTimes(stores lap times of every lap of every driver in every race), pitStops(stores pit stops information like when did the driver pit, how long was the pit stop), qualifying(stores qualifying times of each driver in all 3 qualifyings),races(stores race data like time and location of race), results(stores results of each driver and constructor race wise), seasons(stores external links to data about each season), status(stores what each status Id refers to like status 1 represents driver status as Finished).

IV. Our approach

A. Motivation

Our approach to devise driver rankings based on pure talent control for only the requried variables. Related work done on the the topic of formula one driver analysis all involved finding one common set of rankings between drivers of all decades, We think that this method of comparing all drivers under one model may not be the best approach due to the following reason- Car dynamics and design style have changed majorly over the years, it started with the major factor affecting a car to be the engine and the transmission, to major factors not just revolving around the engine but including factors like car aerodynamics, car setting, hybrid power.There are multiple real world examples to prove these factors like in the 2014 to 2016 the Mercedes developed engine was very competitive and was alone enough to push the teams that used Mercedes engines in their cars like Williams and Force India to the top of the grid.Contrary to the Mercedes engine still being dominant in current seasons but not being able to single handedly allow Williams(currently last) and Force India to top of the grid. We can compare the hybrid power fact by looking at the extraordinary performance of driver Sebastian Vettel in late 2000s when there was no hybrid power and he had to drive against so many greats in F1 and still won 4 world championships to his current performance which is not exactly of the same kind. Other factors which cannot be considered while taking a season wide driver ranking are car safety, rivals the driver has to compete against and not being able to fully estimate the cars performance. Car safety has improved greatly over the years and it gives the drivers a set of training wheels to rely on even when things go south, this was not the case in 1900s as crash could very well be fatal, this could mean the extent to which the drivers would push themselves would be different.Rivals in formula one are a major factor which determine the drivers outcome, It is said that Alain Prost and Ayrton Senna could never have achieved what they achieved without having each other to comepete against in their time, It is not necessary that rivals will always have a good impact, rivals could have a negative impact on other drivers too.

B. Our solution

We perform join operations on driver,constructor, results, status and circuit tables to get a combined table with all important attributes needed to perform analysis. To negate the effects caused by the factors mentioned under Motivation we split the entire history of formula 1 data into 7 decades (1950-60,1960-70,1970-80,1980-90,1990-2000,2000-2010,2010-2020) These splits are based on looking at the major changes in formula one regulations and it was found that taking 10 years period is optimal since lesser than 10 years would mean having less data to fit the model, more would mean data not exactly belonging to the same distribution. On these 7 splits we train separate linear regression models. The independent variables given to the model are driverRef, team-year, circuitId, statusId and the dependent variable is points scored. driverRef is a string used to uniquely identify a driver. team-year is a string made with the concatenation of constructorRef which uniquely identifies each constructor and the year which is the year in which the race is happening. circuitId is an Id given to each circuit to uniqely identify circuits(tracks). statusId is the status of the driver at the end of a race, whether he was able to finish the race, if not what were the reasons for failure, if the reasons were driver based or just random. All driver related issues for which the driver is to penalized for are taken as one category, all random issues are categorized under one label and driver finishing under one category. The dependent variable that is the points scored by the driver in a race was calculated based on the finishing position of the driver,first place gets 25 points,second gets 18, third 15 and 12,10,8,6,4,2,1 to positions fourth through tenth. For drivers finishing below 10th we use a formula to give fractional points.

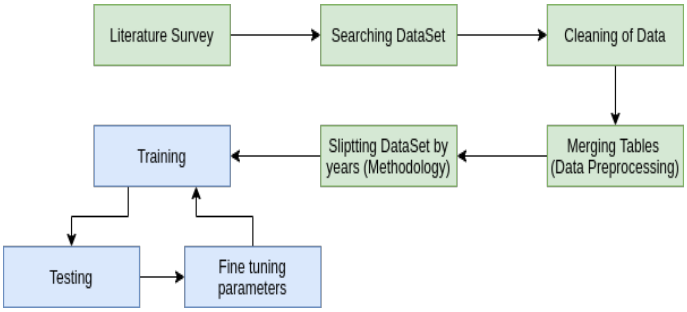

Fig. 1. System Design

C. System Design

Our system design started with collecting the data in the form of 13 tables. These tables were then studied and tables important for our study namely driver, constructor, results, status and circuit were merged into one table. The data was then preprocessed. The preprocessed data was then split 7 ways (decade wise). Individual regression models were trained on each of these seven splits. The fit summary was taken for each model. Results were then given, decade wise.

D. Technologies Used

The entire code was written in python, using Google colaboratory. The modules in python used were Pandas,Numpy,Statsmodels,Scikit to work with dataframes and train models.

E. Model

We use the OLS linear regression model in Statsmodels module for each data split with points scored as the dependent variable, driver reference, team year, status, circuitId as the independent variables. For each of these seven models, the coefficients assigned to the driver reference variable were taken. These coefficients are the true driver skill levels predicted by our models. The driver coefficient with the largest value is the best driver and the one with the lowest is the worst.

V. Observations

The driver rankings were found for each of the seven splits and the results were as follows 1950-1960 Fangio came out on top with a skill rating of 4.66, followed by Beaufort at 3.57 and so on. Fangio who is considered one of the greatest time of all time came out on top. Previously done models [1],[2] gives Fangio the highest ranking thus this model is in accordance and works correctly.

1960-1970 Jim Clark came out on top with a skill rating of 5.77, Jim like Fangio is regarded as one of the best drivers of all time and [1] gives him a rank of 2 just after Fangio.

1970-1980 Rindt,Mclaren,Stewart,Hunt come out in that order with skill ratings 2.73, 2.16, 1.66, 1.12. These names also are very highly regarded by Formula One experts and highly ranked by [1] and [2].

1980-1990 Senna and Prost come out in positions 1 and 2 with skill ratings of 2.21 and 2.01. Senna and Prost are also considered as one of the greatest drivers of all time.They have 4 and 5 world championships under their name.[1] and [2] rank Prost higher than Senna which is different from our model. This is due to the fact that Senna raced in fewer races and still got very high number of wins.

1990-2000 Prost, Micheal Schumacher, Senna were ranked in that particular order with 3.34,2.27 and 2.08. Micheal Schumacher has the world record for most number of world titles(7) tied by Lewis Hamilton and is very highly regarded by all F1 experts. Reason for him not coming out in first place is due to him entering more races than Prost. This result is in accordance with [1] and [2].

2000-2010 Micheal Schumacher comes out on top followed by alonso and lewis Hamilton with skill ratings 4.06, 3.62, 3.59. All these drivers are very high regarded and the ratings are in accordance with [1] and [2].

2010-2020 Alonso, Hamilton, Max Verstappen come out on top with ratings of 5.0, 4.0, 3.9. Alonso and Hamilton are world champions, the reason for Max Verstappen being rated so high is that he is regarded as one of the best drivers(experts consider him to be almost as good as hamilton, our model shows this relationship very closely).Another reason for all 3 of them being ranked at the top is they always outperform their teammates.

VI. Conclusions

This project is the first one which predicts driver rankings on slices of data ie different rankings for different decades. The driver true skill level can be generated by using points scored by drivers in each race and finding out how much of that was because of the driver itself (separating constructor effect,circuit effect from the points scored) Our linear regression models produce the driver rankings which can be used by Formula one constructors to make hires for next seasons.The formula one brand also make such predictions from time to time to keep fans interested and keep the revenue stable.

References

[1] Andrew J. K. Phillips, 2014 ”Uncovering Formula One driver performances from 1950 to 2013 by adjusting for team and competition effects”

[2] Reiner Eichenberger, David Stadelmann, 2009 ”Who Is The Best Formula 1 Driver? An Economic Approach to Evaluating Talent”

[3] Andrew Bell, James Smith, Clive E Sabel, Kelvyn Jones, 2016 ”Formula for success: Multilevel modelling of Formula One Driver and Constructor performance, 1950-2014”

[4] Tobias Lampprecht, David Salb, Marek Mause, Huub van de Wetering, Michael Burch, Uwe Kloos 2019 ”Visual Analysis of Formula One Races”

[5] Stadelmann, D. 2006 ”Who Is The Best Formula 1 Driver? An Econometric Analysis”